PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-189502
(43)Date of publication of application : 05.07.2002

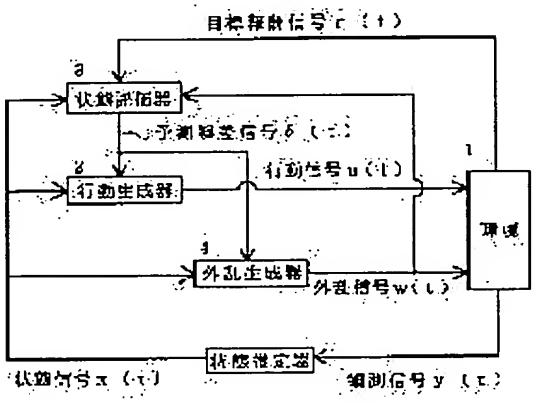
(51)Int.Cl. G05B 13/02
G05B 13/04
G06N 3/00

(21)Application number : 2000-386265
(22)Date of filing : 20.12.2000

(71)Applicant : JAPAN SCIENCE & TECHNOLOGY CORP
ADVANCED TELECOMMUNICATION RESEARCH INSTITUTE INTERNATIONAL
(72)Inventor : MORIMOTO ATSUSHI
DOTANI KENJI

(54) ROBUST REINFORCE LEARNING SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a method for learning robust control which is relatively resistant to the fluctuation of an environment capable of complying even with the case that an environment model is unknown and to provide a robust controller.
SOLUTION: This system is provided with an action generator (2) having a learning function and also outputting an action signal u (t) to an environment (1), a disturbance generator (4) having a learning function and also outputting a disturbance signal w (t) to an environment, and a state evaluator (3) generating an evaluation signal q (t) being a reward signal obtained by adding a reward corresponding to the level of achievement of a target with a reward corresponding to resistance to disturbance from the disturbance generator, and predicting the expected value of the sum of the evaluation signals to be obtained from the present state x (t) to the future, and generating the prediction error signal. Then, the action generator learns to maximize the expected value of the sum of the evaluation signals to be obtained from the present state to the future, while the disturbance generator learns to minimize the expected value of the sum of the valuation signals.



LEGAL STATUS

[Date of request for examination] 20.12.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3465236

[Date of registration] 29.08.2003

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－189502
（P2002－189502A）
(43)公開日　平成14年７月５日(2002.7.5)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０５Ｂ　13/02		Ｇ０５Ｂ　13/02	Ｔ	５Ｈ００４
			Ｊ	
13/04		13/04		
Ｇ０６Ｎ　3/00	５５０	Ｇ０６Ｎ　3/00	５５０Ｅ	

審査請求　有　　請求項の数９　ＯＬ　（全 13 頁）

(21)出願番号　特願2000－386265(P2000－386265)

(22)出願日　平成12年12月20日(2000.12.20)

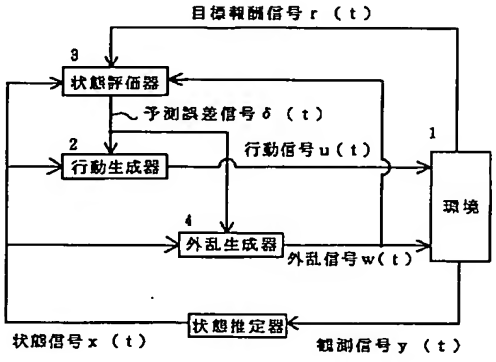
特許法第30条第１項適用申請有り

(71)出願人　396020800
科学技術振興事業団
埼玉県川口市本町４丁目１番８号
(71)出願人　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(72)発明者　森本　淳
奈良県生駒市高山町8916－５　学生宿舎２－102
(74)代理人　100099265
弁理士　長瀬　成城

最終頁に続く

(54)【発明の名称】　ロバスト強化学習方式

(57)【要約】
【課題】環境モデルが未知の場合にも対応することができるとともに、環境の変動に比較的強いロバスト制御の学習方法およびロバスト制御器を提供する。
【解決手段】環境（１）に行動信号ｕ（ｔ）を出力するとともに学習機能を具備する行動生成器（２）と、環境に外乱信号ｗ（ｔ）を出力するとともに学習機能を具備する外乱生成器（４）と、目標の達成度に応じた報酬に、前記外乱生成器からの外乱に耐えうることに応じた報酬を加味した報酬信号である評価信号ｑ（ｔ）を生成し、現在の状態ｘ（ｔ）から将来に向けて得られる評価信号の和の期待値を予測し、その予測誤差信号を生成する状態評価器（３）とを備え、現在の状態から将来に向けて得られる評価信号の和の期待値を最大化するべく行動生成器は学習し、一方、外乱生成器は前記評価信号の和の期待値を最小化すべく学習する。



1

【特許請求の範囲】

【請求項１】制御対象あるいは環境に行動信号を出力するとともに学習機能を具備する行動生成器、および制御対象あるいは環境に外乱信号を出力するとともに学習機能を具備する外乱生成器を備え、

目標の達成度に応じた報酬に、前記外乱生成器からの外乱に耐えうることに応じた報酬を加味した報酬信号である評価信号を生成し、現在の状態から将来に向けて得られる評価信号の荷重和の期待値を最大化（または最小化）するべく行動生成器は学習し、一方、外乱生成器は前記評価信号の和の期待値を最小化（または最大化）すべく学習することを特徴とするロバスト強化学習方式。

【請求項２】前記学習方式において、現在の状態から将来に向けて得られる評価信号の和の期待値を予測する状態評価器を備え、その予測誤差信号を、状態評価器、行動生成器、および外乱生成器の少なくとも１個の学習に用いることを特徴とするロバスト強化学習方式。

【請求項３】前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、入出力関係を示す参照テーブルを具備していることを特徴とする請求項１または２に記載のロバスト強化学習方式。

【請求項４】前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、線形モデルまたは多項式モデルを具備していることを特徴とする請求項１または２に記載のロバスト強化学習方式。

【請求項５】前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、多層神経回路網を具備していることを特徴とする請求項１または２に記載のロバスト強化学習方式。

【請求項６】請求項１または請求項２の方式により、予め学習された前記状態評価器と行動生成器または行動生成器のみを用いた制御方式。

【請求項７】請求項１または請求項２の方式を計算機シミュレーションによって実現される環境モデルに適用し、それによって学習された前記状態評価器と行動生成器または行動生成器のみを実環境に適用することを特徴とする請求項６に記載の制御方式。

【請求項８】前記状態評価器または行動生成器の少なくとも一方は、関数近似手段として、入出力関係を示す参照テーブルを具備していることを特徴とする請求項６または７に記載のロバスト制御方式を用いたロバスト制御器。

【請求項９】前記状態評価器または行動生成器の少なくとも一方は、関数近似手段として、線形モデル、多項式

2

モデルまたは多層神経回路網を具備していることを特徴とする請求項６または７に記載のロバスト制御方式を用いたロバスト制御器。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、強化学習において環境の不確かさに対してロバストな行動則を学習する方式に関するものである。さらに、詳しくは、ロボット、自動車、航空機などの物理的システムの制御、また、広く人間に代わって情報検索、ユーザー応答、資源割当、市場取引などを行うコンピュータプログラムなどにおいて、環境の様々な外乱や、想定していたモデルと現実との環境のズレに対応することができる強化学習方式および強化学習された制御器である。なお、この明細書においては、特許明細書では使用不可能な文字に対応するために、下記の様にして対処している。

【外１】

Ｈ∞はＨ∞で表している。

$\int_t^\infty$ は∫ｔ∞で表している。

【０００２】

【従来の技術】従来の強化学習方式について説明する。図２は従来の学習方式に用いる回路のブロック図である。図３は制御対象と制御器とのフィードバックのブロック図であり、Ｈ∞制御理論を説明するための図である。環境１は制御対象をはじめ、行動の対象となるシステム一般を表し、状態信号ｘ(t) を行動生成器２および状態評価器３に出力している。行動生成器２は状態信号ｘ(t) が入力されると行動信号ｕ(t) を環境１に出力している。また、環境１から報酬信号ｒ(t) が状態評価器３に入力される。状態評価器３は、目標の達成度に応じた報酬ｒ(t) が入力されると、現在の状態ｘ(t) から将来に向けて得られる、報酬ｒ(t) の荷重和の期待値すなわち、評価関数Ｖ（ｘ(t) ）を予測し、その予測値を用いて予測誤差信号δ(t) を生成し、行動生成器２に出力する。行動生成器２は、状態評価器３から予測誤差信号δ(t) が入力されると、現在の状態ｘ(t) から将来に向けて得られる、報酬ｒ(t) の荷重和の期待値すなわち、評価関数Ｖ（ｘ(t) ）が最大となる様に学習し、その入出力の関係を変更する。ただし、前記評価関数Ｖ（ｘ(t)）は、連続の場合、

【数１】

3 4

$$V(x(t))=E\left[\int_t^{\infty} e^{-\frac{s-t}{\tau}} r(x(s),u(s))ds\right]$$

（式１）

（τは評価の時定数）、離散系の場合

$$V(x_T)=E\left[\sum_{s=t}^{\infty} \alpha^{s-t} r_s\right]$$

（αは評価の減衰率）

と定義される。

この学習方式では、ある環境のもとで最適な行動が学習されるが、異なる環境では動作は保証されていない。また、異なる環境に適応するためには再学習を行う必要があり、その再学習の時間が新たに必要となる。

【０００３】次に、従来のＨ∞制御について説明する。図３において、制御対象Ｇから観測信号ｙ(t) が制御器Ｋに入力される。制御器Ｋは観測信号ｙ(t) が入力されると、制御信号である行動信号ｕ(t) を制御対象Ｇに出力する。また、制御対象Ｇには外乱信号ｗ(t) および行動信号ｕ(t) が入力され、これらの信号が入力されると、評価用信号ｚ(t) および観測信号ｙ(t) を出力する。なお、外乱の影響を評価するための評価用信号ｚ(t) と、制御対象Ｇを観測して制御器Ｋに入力するフィードバック信号である観測信号ｙ(t) とは、同じにする＊ことも可能であるが、異ならしめることも可能である。そして、ロバスト制御の代表的な定式化であるＨ∞制御問題の要請は、図３に図示するフィードバック系で未知外乱やモデル誤差に起因する外乱信号ｗ(t) による評価用信号ｚ(t) への影響を少なく抑えつつ、出力を安定化する。すなわち、評価用信号ｚ(t) ＝０に近づけることである。具体的には、Ｈ∞ノルムによりシステムの外乱に対する感度を測り、ロバスト性の基準値γ以下となるような制御器Ｋの設計を行う。ノルムとは、ある種の大きさの指標であり、外乱信号ｗ(t) から評価用信号ｚ(t) への伝達関数行列をT_{zw}としたとき、そのＨ∞ノルム、$\|T_{zw}\|_\infty$は次の（式２）のように定義される。

【数２】

$$\|T_{zw}\|_\infty = \sup_w \frac{\|z\|_2}{\|w\|_2} < \gamma \qquad （式２）$$

ただし、supｗは外乱信号ｗ(t) に関する上限を表し、外乱信号ｗ(t) を変化させたときに、$\|z\|_2 / \|w\|_2$ が supｗ（$\|z\|_2 / \|w\|_2$）より大きくならないことを示している。また、$\|z\|_2$ および $\|w\|_2$ はそ※れぞれ評価用信号ｚ(t) および外乱信号ｗ(t) のＬ₂ノルムであり、次の（式３）および（式４）で定義される。

【数３】

$$\|z\|_2 = \left(\int_0^{\infty} z^T(t) z(t)\, dt\right)^{\frac{1}{2}} \qquad （式３）$$

【数４】

$$\|w\|_2 = \left(\int_0^{\infty} w^T(t) w(t)\, dt\right)^{\frac{1}{2}} \qquad （式４）$$

【０００４】ここで、評価関数Ｖを次の（式５）の様に定義する。★【数５】★

$$V=\int_0^{\infty} (-z^T(t) z(t) + \gamma^2 w^T(t) w(t))\, dt$$

（式５）

これを行動信号ｕ(t) に関しては最大化し、外乱信号ｗ(t) に対しては最小化する問題を考える。その結果、Ｖ≧０を満たす解を得られれば、（式２）の条件のもとで、評価用信号ｚ(t) の安定化が実現できる。

【０００５】強化学習の課題のうち、予め与えられた目標点あるいは目標軌道への近さを報酬信号とするもの

は、学習制御の課題と考えることができる。前記ロバスト制御の代表的な方法であるＨ∞制御は、システムの外乱による影響の受けやすさをＨ∞ノルムで評価し、フィードバック系のＨ∞ノルムを一定以下に抑える制御器を設計することにより、外乱やモデル誤差に対するロバスト性を保証するものである。しかし、その制御器の解析的な構成手法は線形システムに対し示されており、非線形システムに対してはある限定されたシステムを除いては、一般に解析的に制御器を構築する方式はない。非線形システムにおいて、多層神経回路網を用いて未知外乱を考慮した状態価値関数を近似しロバスト制御を実現する手法が提案されているが、これらは、制御器の適応可能範囲が線形近似可能な領域付近に限られていたり、学習がオフラインのバッチ学習に限られている。また、これらの学習には環境のモデルを必要としている。さらに、状態を離散化し動的計画法を用いる方式も提案されているが、制御器の構築には、状態を離散化する過程と、オフラインの計算過程とを必要とし、かつ、環境モデルを必要とする。また、ここまでに挙げたロバスト制御器はレギュレータ（目標点を原点とし、その原点に制御対象の状態を持って行く制御）としてのみ機能する。

【０００６】ところで、従来のミニマックス戦略を用いる強化学習方式では、オセロやバックギャモンの学習を行っている。また、戦闘機やミサイルの逃亡追従問題に適用した例もある。一方、本発明では、仮想的な敵を想定して学習を行った後、実際に制御器を適用する際には、その外乱生成器を除いて制御器を構成するという点で異なる。これは、外乱を仮想的な敵とみなすことによってロバストな制御器を強化学習により構築するという新しい発想に依存している。また、後述の（式８－２）で定義される評価関数の予測誤差に重み付けをして学習することで、ロバストな行動則を強化学習によって獲得する手法とは異なり、本発明はＨ∞制御との関連や目的関数が明確である。よって、最悪の外乱を求めているという点において、より確実にロバストな制御器を獲得することができる。

【０００７】そして、後述のように、本発明の実施の形態においては、非線形、オンライン、モデル非依存性の動的最適化手法である強化学習方式に、ロバスト制御における最悪外乱に対する感度の最小化の原理を導入することにより、ロバスト強化学習方式を実現し、その制御課題への適用により、非線形、オンライン、モデル非依存性のロバスト学習制御を実現する。また、環境の近似モデルが既知の場合、あるいは未知の場合でも、環境のモデルを学習することにより、モデル依存の強化学習方式を拡張し適用することにより、効率良く非線形ロバスト制御器をオンラインで構築することができる。

【０００８】この様なことができる利点は、環境が非定常であった場合に、一度学習によって構築した非線形ロバスト制御器や環境モデルを用いれば、ある程度の環境の変動なら再学習することなく対応することができることである。また、従来のロバスト制御は出力をある目標値に収束させるレギュレータ問題に対してのみ定式化されていたが、本発明のロバスト強化学習方式は、任意の評価基準に対して適用可能である。

【０００９】

【発明が解決しようとする課題】この様に、従来の強化学習では、環境の変動が生じた場合には、再学習する必要があり、従来のロバスト制御の方式は環境のモデルが未知の場合には対応することができなかった。さらに、非線形系に適用可能な制御則の具体的方法は与えていない。

【００１０】そこで本発明は、環境モデルが未知の場合にも対応することができるとともに、非線形系に適用可能な環境の変動に強いロバスト強化学習方式およびロバスト制御器を提供することを目的とする。

【００１１】

【課題を解決するための手段】このため、本発明が採用した課題解決手段は、制御対象あるいは環境に行動信号を出力するとともに学習機能を具備する行動生成器、および制御対象あるいは環境に外乱信号を出力するとともに学習機能を具備する外乱生成器を備え、目標の達成度に応じた報酬に、前記外乱生成器からの外乱に耐えうることに応じた報酬を加味した報酬信号である評価信号を生成し、現在の状態から将来に向けて得られる評価信号の荷重和の期待値を最大化（または最小化）するべく行動生成器は学習し、一方、外乱生成器は前記評価信号の荷重和の期待値を最小化（または最大化）すべく学習することを特徴とする強化学習方式であり、前記学習方式において、現在の状態から将来に向けて得られる評価信号の和の期待値を予測する状態評価器を備え、その予測誤差信号を、状態評価器、行動生成器、および外乱生成器の少なくとも１個の学習に用いることを特徴とする強化学習方式であり、前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、入出力関係を示す参照テーブルを具備していることを特徴とする強化学習方式であり、前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、線形モデルまたは多項式モデルを具備していることを特徴とする強化学習方式であり、前記状態評価器、行動生成器および外乱生成器の少なくとも一個は、関数近似手段として、多層神経回路網を具備していることを特徴とする強化学習方式であり、前記方式により、予め学習された前記状態評価器と行動生成器または行動生成器のみを用いた制御方式であり、前記方式を計算機シミュレーションによって実現される環境モデルに適用し、それによって学習された前記状態評価器と行動生成器または行動生成器のみを実環境に適用することを特徴とする制御方式であり、前記状態評価器または行動生成器の少なくとも一方は、関数近似手段として、入出力関

係を示す参照テーブルを具備していることを特徴とするロバスト制御器であり、前記状態評価器または行動生成器の少なくとも一方は、関数近似手段として、線形モデル、多項式モデルまたは多層神経回路網を具備していることを特徴とするロバスト制御器である。

【００１２】

【実施の形態】次に、本発明における強化学習方式およびロバスト制御器の実施の一形態を説明する。図１は本発明の実施の形態のロバスト強化学習方式に用いる回路のブロック図である。図４は第１具体例の説明図で、（ａ）が概略図、（ｂ）が１ｍの長さの振り子の角度変化のグラフである。図５は第１具体例の角度変化のグラフで、（ａ）が０．５ｍの長さの振り子のグラフ、（ｂ）が２ｍの長さの振り子のグラフである。図６は第２具体例の概略図である。図７は第２具体例の位置変化のグラフで、（ａ）が１ｋｇの質量の搬送物を用いた場合のグラフ、（ｂ）が３ｋｇの質量の搬送物を用いた場合のグラフである。

【００１３】本発明では、外乱や環境の変化に対してロバストな強化学習を実現するため、目標の達成度に応じた報酬ｒ（ｔ）に対して、外乱に耐えうることに応じた報酬ｓ（ｔ）を加えた新たな報酬である評価信号ｑ(t)を次のように定義した時、

【００１４】

【数６】

$$q(t) = r(x(t), u(t)) + s(w(t)) \quad (式6)$$

この評価信号ｑ（ｔ）を報酬とした最大最小問題を強化学習方式の枠組みにおいて解く。よって、前述のＨ∞制御の問題設定は、本発明の実現例の一つとなっていることがわかる。以上を考慮した上で、次のようにロバスト＊強化学習の学習方式に用いる回路を図１に図示するように構築する。

【００１５】この図１の説明において、図２の従来の回路と同じ構成要素には同じ符号を付して、その説明は省略する。この図１においては、外乱生成器４が設けられている。そして、環境１からの観測信号ｙ（ｔ）が、状態推定器を介して状態信号ｘ(t)となり、行動生成器２、状態評価器３および外乱生成器４に入力されている。この外乱生成器４は状態信号ｘ(t)が入力されると、外乱信号ｗ(t)を環境１および状態評価器３に出力する。この状態評価器３は、環境１からの目標報酬信号ｒ(t)と、外乱生成器４からの外乱信号ｗ(t)に基づいて生成した外乱報酬信号ｓ(t)とに基づいて、現在の状態ｘ(t)から将来に向けて得られる評価信号ｑ(t)の荷重和の期待値を予測し、その予測値に基づいて予測誤差信号δ(t)を生成し、行動生成器２および外乱生成器４に出力する。この様にして、状態評価器３は、外乱報酬信号ｓ(t)に目標報酬信号ｒ(t)を加算して評価信号ｑ(t)を得て、予測誤差信号δ(t)を生成し出力している。そして、行動生成器２は予測誤差信号δ(t)が入力されると、現在の状態ｘ(t)から将来に向けて得られる上記評価信号ｑ(t)の荷重和の期待値を最大化するべく学習し、その入出力関係を変更する。一方、外乱生成器４は予測誤差信号δ(t)が入力されると、現在の状態ｘ(t)から将来に向けて得られる上記評価信号ｑ(t)の荷重和の期待値を最小化するべく学習し、その入出力関係を変更する。

【００１６】状態評価器３は、（式７）で定義される現在の状態ｘ(t)から将来に向けて得られる評価信号ｑ(t)の期待値Ｖ（ｘ(t)）を予測する。ただし、τは評価の時定数である。

【数７】

$$V(x(t)) = E\left[\int_t^{\infty} e^{-\frac{s-t}{\tau}} q(s)\, ds\right] \quad (式7)$$

行動生成器２および外乱生成器４は、この期待値Ｖ（ｘ(t)）がそれぞれ、最大化、最小化される様に行動信号ｕ(t)、外乱信号ｗ(t)を学習する。なお、行動生成器２、状態評価器３および外乱生成器４としては、参照テーブル、線形モデル、多項式モデル、多層神経回路網などを用いることができる。

【００１７】ロバスト強化学習を行う時点においては、状態評価器３、行動生成器２および外乱生成器４は同時に作動させるが、実際に学習した行動則を制御対象または環境に用いる段階においては、状態評価器３および行動生成器２、或いは、行動生成器２のみを用いて動作させる。この行動生成器２には、観測信号ｙ(t)として状態信号ｘ(t)が直接得られる場合はそれを用いるが、一般にはオブザーバ、カルマンフィルタなどにより状態信号ｘ(t)を推定し入力として用い、また、学習時には、環境１はモデルでも、実際の環境でも可能である。そして、実際の環境の場合には、行動信号ｕ(t)および外乱信号ｗ(t)は、アクチュエータなどの駆動源や、低レベルの制御プログラムへの指令などの作動手段を介して環境１に入力される。一方、報酬信号ｒ(t)や状態信号ｘ(t)は、センサーなどの検知手段を介して環境１から出力される。

【００１８】そして、状態評価器３は、評価関数Ｖ（ｘ(t)）のパラメータ$v = \{v_1, v_2, \cdots, v_l, \cdots\}$を持つ近似器Ｖ（ｘ(t)；ｖ）として実現され、その手段としては、前述の参照テーブル、線形モデル、多項式モデルおよび多層神経回路網を用いることができる。

【００１９】この様にして、状態評価器３は、環境１か

9

ら目標報酬信号ｒ(t) を得る手段と、外乱生成器４から外乱信号ｗ(t) を得る手段と、目標報酬に外乱報酬を加味した評価信号ｑ(t) を得る手段と、現在の状態から将来に向けて得られる評価信号ｑ(t) の和の期待値を予測し、予測誤差信号δ(t) を生成する手段とを有している。また、行動生成器２は、環境１から状態信号ｘ(t) を得る手段と、状態評価器３から予測誤差信号δ(t) を得る手段と、環境１に行動信号ｕ(t) を出力する手段と、現在の状態から将来に向けて得られる評価信号ｑ(t) の和の期待値が最大化する様に学習する手段とを有している。そして、外乱生成器４は、環境１から状態信号ｘ(t) を得る手段と、状態評価器３から予測誤差信号δ(t) を得る手段と、環境１に外乱信号ｗ(t) を出力する＊

10

＊る手段と、現在の状態から将来に向けて得られる評価信号ｑ(t) の和の期待値が最小化する様に学習する手段とを有している。

【００２０】以降、離散系での評価関数の学習、連続系での評価関数の学習、離散系での行動決定方法、連続系での行動決定方法の順に示す。離散系での評価関数の学習：次の様な確率分布Ｐにしたがう動的制御対象を考える。

Ｐ（$x_{\tau+1}$ | x_τ , u_τ , w_τ ）

ただし、x_τ は状態変数、u_τ は制御入力、w_τ は外乱入力を表す。このとき、求めるべき評価関数V_τ は次の式で表される。

【数８】

$$V_\tau = \sum_{S=T}^{\infty} \alpha^{S-T} q_s \qquad (式８)$$

ただし、q_τ はただちに得られる評価信号、α（０≦α≦１）は評価の減衰率を表す。そこで、状態評価値の予※

※測誤差δ_τ は次式のように表される。

$$\delta_\tau = q_\tau + \alpha V_{\tau+1} - V_\tau \qquad (式８－２)$$

【００２１】よって、この予測誤差δ_τ と、次の（式９）で表される各パラメータの寄与度の履歴$e_{i\tau}$を用い★

【数９】

$$e_{i\tau} = \sum_{n=0}^{T} (\alpha\lambda)^{T-n} \frac{\partial V_n}{\partial v_i} \qquad (式９)$$

パラメータの更新量Δv_i は、次式のように表される。

$\Delta v_i = \eta \delta_\tau e_{i\tau}$

ただし、λはパラメータの寄与度の履歴の減衰率を、η☆は学習率を表す。また、各パラメータの寄与度の履歴$e_{i\tau}$は（式９）の定義より次式を用いて更新される。

【数１０】

$$e_{i\tau} = \alpha\lambda e_{i(\tau-1)} + \frac{\partial V_\tau}{\partial v_i} \qquad (式１０)$$

【００２２】連続系での評価関数の学習：次の様な動的制御対象を考える（状態変数ｘ(t)の時間変化ｄｘ／ｄｔを、状態変数ｘ(t)、制御入力ｕ(t)、外乱入力ｗ(t)、ノイズ入力ｎ（ｔ）の関数として考える）。◆

◆ｄｘ／ｄｔ＝ｆ（ｘ(t),ｕ(t),ｗ(t))＋ｎ（ｔ）

ただし、このとき、求めるべき評価関数Ｖ(t) は次式で表される。

【数１１】

$$V(t) = E\left[\int_t^{\infty} e^{-\frac{s-t}{\tau}} q(s)\, ds\right] \qquad (式１１)$$

ただし、ｑ(t) はただちに得られる報酬であり、τは評価値の時定数である。よって、状態評価値の予測誤差δ(t) は次式のように表される。

δ(t) ＝ｑ(t) －（１／τ）×Ｖ(t) ＋ｄＶ(t) ／ｄｔ＊

＊【００２３】ここで得られる状態評価値の予測誤差δ(t) と、次式で表される各パラメータの寄与度の履歴を用いて、

【数１２】

$$e_i(t) = \int_0^t e^{-\frac{t-s}{k}} \frac{\partial V(s)}{\partial v_i} ds \qquad (式１２)$$

ただし、ｋはパラメータの寄与度の履歴の時定数である。状態評価器のパラメータの更新量ｄｖ$_i$／ｄｔ（連続系ではパラメータｖ$_i$の時間微分で表される）は次式のように表される。

ｄｖ$_i$／ｄｔ＝ηδ(t) ｅ$_i$ (t)

ただし、ηは学習率を表す。

【００２４】また、各パラメータの寄与度の履歴ｅ$_i$ (t) の更新量ｄｅ$_i$ (t) ／ｄｔは、（式１２）の定義

11 12

により次式を用いて更新される。 ＊ ＊【数１３】

$$\frac{de_i(t)}{dt} = -\frac{1}{k} e_i(t) + \frac{\partial V(t)}{\partial v_i} \quad (式１３)$$

【００２５】離散系での行動決定方法：
（モデル非依存の場合）次式で示す確率分布Ｐｒに従っ※て、状態ｓにおける行動ａを決定する。

【数１４】

$$P_r(a_T = a \mid s_T = s) = \frac{e^{\beta A(s,a)}}{\sum_b e^{\beta A(s,b)}} \quad (式１４)$$

ただし、Ａ(s,a) は行動決定のためのパラメータであり、状態ｓにおける行動ａの取りやすさを表している。また、βは行動のランダムさを表すパラメータである。この時、行動生成器のパラメータ更新量ΔＡ$_{a}$（s_T , a_{aT}）と、外乱生成器のパラメータ更新量ΔＡ$_{e}$（s_T , a_{eT}）は（式８－２）の予測誤差δ_Tを用いて次式でそれぞれ表される。

$\Delta A_a(s_T, a_{aT}) = \eta^{\wedge}_{a}\, \delta_T$

$\Delta A_e(s_T, a_{eT}) = -\eta^{\wedge}_{e}\, \eta_A\, \delta_T$ ★

★ただし、$\eta^{\wedge}_{a}$, $\eta^{\wedge}_{e}$ は学習率を表す。また、a_{aT}, a_{eT}はそれぞれ、時刻Ｔにおける行動生成器と外乱生成器の行動を表す。

【００２６】モデル非依存の学習方式として、行動価値関数を学習することによって、ロバスト強化学習を実現することができる。つまり、次式で表される行動価値関数の予測誤差δ_Tを用いて、

【数１５】

$$\delta_T = q_T + \alpha \max_{a_a} \min_{a_e} Q(s_{T+1}, a_a, a_e) - Q(s_T, a_{aT}, a_{eT}) \quad (式１５)$$

行動価値関数の更新量ΔＱ(s_T , a_{aT}, a_{eT}) は次式のようになる。

$\Delta Q(s_T, a_{aT}, a_{eT}) = \eta_0\, \delta_T$

ただし、η_0は学習率、αは評価の減衰率である。 ☆

☆【００２７】次式で示す確率分布Ｐｒ$_{e}$に従って、状態ｓにおいて、外乱生成器の行動a_{eT+1}をすべての行動生成器の行動a_{aT}に関して決定する。

【数１６】

$$P_{re}(a_{eT+1} = a_e \mid a_{aT} = a_{aj}, s_T = s) = \frac{e^{-\beta Q(s, a_{aj}, a_e)}}{\sum_{b_e} e^{-\beta Q(s, a_{aj}, b_e)}} \quad (式１６)$$

この場合、確率分布Ｐｒ$_{e}$に従うことで、小さい行動価値を持つ外乱生成器の行動ａ$_{e}$を高い確率で選択することになる。これによって、目的とする課題の達成にとって外乱生成器が最悪の外乱を生成するようになる。ただし、βは行動のランダムさを表すパラメータである。

【００２８】次に、次式で表される確率分布Ｐｒ$_{a}$に従って、状態ｓにおいて、行動生成器の行動a_{aT}をすでに決定した外乱生成器のa_{eT}に対する行動a_{eT+1}を用いて決定する。

【数１７】

13 14

$$P_{ru}(a_{u\tau}=a_{u,i}\mid s_\tau=s)=\frac{e^{\beta Q(s,a_{u,i},a_{w\tau,j})}}{\sum_{b_{u,i}} e^{\beta Q(s,b_{u,i},a_{w\tau,j})}}$$

（式１７）

ただし、行動生成器が行動$a_{u,i}=a_{u,i}$を選択した時、外乱生成器は行動$a_{w,\tau}=a_{w,\tau,i}$を選択する。この場合、確率分布P_{ru}に従うことで、大きい行動価値を持つ行動生成器の行動$a_{u,i}$を高い確率で選択することになる。これによって、目的とする課題の達成にとって行動生成器が最高の行動出力を生成するようになる。ただし、βは行動のランダムさを表すパラメータである。

＊【００２９】離散系での行動決定方法：
（モデル依存の場合）状態X_τにおいて、行動生成器の行動がｕで、外乱生成器の行動がｗであり、その結果状態$X_{\tau,1}$にたどり着いたとする。そのときに得られる評価信号$q_{\tau,1}$とすれば、環境のモデルを用いて、行動生成器の行動u_τは

＊【数１８】

$$u_\tau=\mathrm{argmax}_u \sum_{X_{\tau+1}} P(X_{\tau+1}\mid X_\tau, u, w)\,[q_{\tau+1}+\alpha V(X_{\tau+1})]$$

外乱生成器の行動w_τは

$$w_\tau=\mathrm{argmin}_w \sum_{X_{\tau+1}} P(X_{\tau+1}\mid X_\tau, u, w)\,[q_{\tau+1}+\alpha V(X_{\tau+1})]$$

となる。

ただし、αは評価の減衰率、$P(x_{\tau,1}\mid x_\tau, u, w)$は状態$x_\tau$において行動生成器が行動ｕを出力し外乱生成器が外乱ｗを出力した時、状態$x_{\tau,1}$に遷移する確率。

【００３０】連続系での行動決定方法：
（モデル非依存の場合）行動生成器の行動をｕ（ｔ），外乱生成器の行動をｗ（ｔ）とすると、それぞれ、

※$u(t)=A(x(t);v^{Au})+n_u(t)$
$w(t)=A(x(t);v^{Aw})+n_w(t)$

※のように表される。ただし、$n_u(t)$，$n_w(t)$は探索のためのノイズ入力を表す。それぞれの行動は、パラメータ$v^A=\{v_1^A, v_2^A, \cdots, v_i^A \cdots\}$を持つ近似器$A(x(t);v^A)$として実現され、その手段としては、線形モデル、多項式モデルおよび多層神経回路網などを用いることができる。また、それぞれのパラメータは、前述の予測誤差信号$\delta(t)$を用いて以下のように更新する。

【数１９】

$$\frac{dv_i^{Au}}{dt}=\eta_u^A\delta(t)n_u(t)\frac{\partial A(x(t);v^{Au})}{\partial v_j^{Au}}$$

$$\frac{dv_i^{Aw}}{dt}=-\eta_w^A\delta(t)n_w(t)\frac{\partial A(x(t);v^{Aw})}{\partial v_i^{Aw}}$$

ただし、η_u^A，η_w^Aは学習率である。

（モデル依存の場合）環境のモデルを用いることが可能な場合は、状態評価器の勾配を用いて、モデル非依存性

の場合に比べて効率的に学習を行うことができる。ここで、環境モデルと報酬モデルを次式で表す。

$$\frac{dx}{dt} = f(x) + g_1(x)w + g_2(x)u \quad (式18-1)$$

$$r(x, u) = Q(u) - u^T R(x) u \quad (式18-2)$$

$$s(w) = \gamma^2 w^T w \quad (式18-3)$$

すると、評価関数の勾配と、環境モデルから得られる入力ゲイン$g_1(x)$，$g_2(x)$を用いて、行動生成器の最適出力u_{*}と、外乱生成器の最適出力w_{*}はそれぞれ次※

※式で表される。

【数21】

$$評価関数の勾配：\frac{\partial V}{\partial x} \quad (式19-1)$$

$$u_{*} = \frac{1}{2} R(x)^{-1} g_2^T(x) \left(\frac{\partial V}{\partial x}\right)^T \quad (式19-2)$$

$$w_{*} = -\frac{1}{2\gamma^2} g_1^T(x) \left(\frac{\partial V}{\partial x}\right)^T \quad (式19-3)$$

入力ゲイン$g_2(x)$，$g_1(x)$は必ずしも既知ではなくても、状態評価の学習と同時に環境モデルを学習することによって求めることができる。

【００３１】具体例１：単振り子の振り上げ

図４の様な単振り子の制御にロバスト強化学習を適用し、学習された制御器を用いて単振り子の振り上げを行った例を示す。単振り子は質量ｍ＝１〔ｋｇ〕、長さＬ〔ｍ〕で、状態変数はｘ＝（θ，ｄθ／ｄｔ）であり、振り子の角度と角速度で表す。制御指令ｕ＝Ｔは振り子の回転軸中心での駆動トルクである。したがって、（式１８－１，式１８－２，式１８－３）との対応を考えると、振り子の運動方程式を構成するそれぞれの関数は以下のように与えられる。

【数22】

$$f(x) = \begin{pmatrix} \dot{\theta} \\ \frac{g}{L}\sin(\theta) - \mu\dot{\theta} \end{pmatrix}$$

$$g_1(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$g_2(x) = \begin{pmatrix} 0 \\ \frac{1}{mL^2} \end{pmatrix}$$

$$Q(x) = \cos(\theta) - 1$$

$$R(x) = 0.08$$

$$\gamma = 0.45$$

$\dot{\theta}$ は、文章中ではｄθ／ｄｔで表示する。

よって、評価信号は以下の関数で表される。

【００３２】

$$q(t) = \cos(\theta) - 1 - 0.08u^2 + \gamma^2 w^2$$

状態変数がｘ＝（θ，ｄθ／ｄｔ）であり、学習時の振り子の質量がｍ＝１〔ｋｇ〕、長さＬ＝１〔ｍ〕であることから、（式１９－１、式１９－２、式１９－３）より、行動生成器および外乱生成器は下記のごとくなる。

【数23】

17

$$u_{op} = 6.25 \frac{\partial V}{\partial \dot{\theta}}$$

$$w_{op} = -\frac{1}{0.405} \frac{\partial V}{\partial \dot{\theta}}$$

$\frac{\partial V}{\partial \dot{\theta}}$ は

学習によって獲得される評価関数Ｖ（ｘ）を

$\dot{\theta}$ で偏微分することで得られる。

ここで、状態変数は連続であるので、状態評価器の関数近似手段として多層神経回路網を用いた。

【００３３】このようにして、学習した行動生成器を、非線形ロバスト制御器として採用し、長さＬ＝0.5,1.0,2.0〔ｍ〕の３種の異なる長さを持つシステムに適用した。また、従来の強化学習を用いて学習した制御器に対しても同様の実験を行った。その結果を以下に示す。

【００３４】図４（ｂ）および図５において、実線で図示するように、全ての環境において、非線形ロバスト制御器は単振り子の振り上げに成功している。ただし、グラフの縦軸は振り子の回転角を、横軸は時間を表してい＊

18

＊る。一点鎖線は振り上がった状態を示している。そして、実線は、一点鎖線で示す直線に収束しているため、振り上げに成功していることが分かる。

【００３５】一方、破線で示すように、通常の強化学習で学習した従来の制御器は、学習時に用いた環境と同一の環境以外では振り子を振り上げることができていない。図４（ｂ）に示す様に、学習時の環境（振り子の長さＬ＝1.0[m])で振り上げを行うと、振り上げ軌道が一点鎖線に収束していることが分かるが、図５に示したように、学習時の環境以外（振り子の長さＬ＝0.5[m],2.0[m])の環境下で振り上げを行うと、振り上げ軌道は一点鎖線に収束しておらず、振り上げができていないことが分かる。

【００３６】具体例２：非線形力場における荷物の搬送

ここでは、図６の様な直動アクチュエータ１１に搬送物１２を載せて運搬することを考える。ただし、勾配のために制御対象に非線形性があり、また、アクチュエータ１１を小型化するために、大きな出力が出ないような状況を想定する。制御対象の運動方程式は、勾配をθ、摩擦係数μ＝0.01，重力加速度ｇ＝9.8[m/s^2]とすると次式で表される。

【数２４】

$$da/dt = \dot{a}$$

$$d\dot{a}/dt = (1/(M+m))((F-(M+m)\,g\,\sin\theta)\cos\theta - \mu\dot{a})$$

なお、文章中は$\dot{a}$はda/dt で表示する。

ただし、状態変数はｘ＝（ａ，da/dt)であり、ａはアクチュエータ１１の水平方向の位置を、 da/dtは速度を表す。また、Ｆはアクチュエータ１１が与える力であり、Ｍは搬送物１２の質量、ｍはアクチュエータ１１の質量である。そして、勾配θは、水平位置がａの場合には、θ＝ arctan(cos(πａ))となる。

【００３７】したがって、（式１８－１、式１８－２、式１８－３）との対応を考えると、運搬用アクチュエータ１１の運動方程式を構成するそれぞれの関数は以下のように与えられる。

【数２５】

$$f(x)=\begin{pmatrix}\dot{a}\\ -g\sin\theta\cos\theta-\frac{1}{M+m}\mu\dot{a}\end{pmatrix}$$

$$g_1(x)=\begin{pmatrix}0\\ 1\end{pmatrix}$$

$$g_2(x)=\begin{pmatrix}0\\ \frac{\cos\theta}{M+m}\end{pmatrix}$$

$$Q(x)=\begin{cases}1.0 & (\text{if } 0.4\leq x\leq 0.6)\\ -0.5 & (\text{otherwise})\end{cases}$$

$$R(x)=0.02$$

$$\gamma=0.2$$

ここで、また、（式６）に対応する報酬は以下の関数を＊＊用いた。

$$q(t)=1.0-0.02u^2+\gamma^2 w^2 \quad (\text{if } 0.4\leq a\leq 0.6)$$

$$=-0.5-0.02u^2+\gamma^2 w^2 \quad (\text{otherwise})$$

【００３８】状態変数がx＝（a，da/dt)であり、学習時のアクチュエータ１１の質量がm＝１〔ｋｇ〕、搬送物１２の質量M＝１〔ｋｇ〕であることから、（式１９－２）（式１９－３）より、行動生成器および外乱生成器は下記のごとくなる。

【数２６】

$$u_{op}=12.5\cos\theta\frac{\partial V}{\partial\dot{a}}$$

$$w_{op}=-\frac{1}{0.08}\frac{\partial V}{\partial\dot{a}}$$

$\frac{\partial V}{\partial\dot{a}}$ は

学習によって獲得される評価関数Ｖ（ｘ）を

$\dot{a}$ で偏微分することで得られる。

ここで、状態変数は連続であるので、状態評価器の関数近似手段として多層神経回路網を用いた。

【００３９】そして、ある目標地点（図７において一点鎖線で図示する）に移動させることを学習した。なお、図７では、縦軸にアクチュエータ１１の位置、横軸に時間を取っている。

【００４０】このようにして、学習した行動生成器を、非線形ロバスト制御器として採用し、学習時と同じ質量（M＝１〔kg〕）の搬送物１２を載せた場合と、学習時よりも重い搬送物１２（M＝３〔kg〕）を載せた場合とで、１５〔m〕離れた地点から目標地点まで搬送するシミュレーション実験を行った結果を図７（ａ）および図７（ｂ）に実線でそれぞれ示した。図７（ａ）および（ｂ）の両方の実線は、アクチュエータ１１の軌道が目標地点を示す一点鎖線に収束していることから、搬送物１２の質量が、M＝１〔kg〕，M＝３〔kg〕の両方の場合で搬送を行うことができることが分かる。

【００４１】一方、通常の強化学習で学習した従来の制御器に対しても、同様の実験を行い、その結果を図７に破線で図示した。図７（ａ）の破線で図示したように、搬送物１２の質量がM＝１〔kg〕の場合には、アクチュエータ１１の軌道が、目標地点をしめす一点鎖線に収束していることから、目標地点への搬送に成功していることが分かる。しかし、図７（ｂ）の破線で図示したように、搬送物１２の質量がM＝３〔kg〕の場合には、アクチュエータ１１の軌道が、目標地点をしめす一点鎖線に収束していないことから、目標地点への搬送ができていないことが分かる。この様に、本発明のロバスト強化学習方式を用いて獲得した制御器は、搬送物１２の質量にばらつきがある場合でも、ある程度の範囲内で対応することができる。

【００４２】以上、本発明の実施の形態について説明したが、本発明の趣旨の範囲内で種々の形態を実施することが可能である。

【００４３】

【発明の効果】以上述べた如く、本発明によれば、目標の達成度に応じた報酬に、外乱に耐えうることに応じた報酬を加味した報酬信号である評価信号を生成し、現在の状態から将来に向けて得られる評価信号の和の期待値を最大化するべく行動生成器は学習し、一方、外乱生成器は前記評価信号の和の期待値を最小化すべく学習するので、環境モデルが未知の場合にも対応することができるとともに、環境の変動に強くなることができる。さら

に、非線形の制御対象あるいは環境にも適用可能である。

【図面の簡単な説明】

【図１】本発明の実施の形態のロバスト強化学習方式に用いる回路のブロック図である。

【図２】従来の強化学習方式に用いる回路のブロック図である。

【図３】H∞制御理論を説明するための制御対象と制御器との制御系のブロック図である。

【図４】第１具体例の説明図で、（ａ）が概略図、（ｂ）が１ｍの長さの振り子の角度変化のグラフである。

【図５】第１具体例の角度変化のグラフで、（ａ）が０．５ｍの長さの振り子を制御対象として用いた場合のグラフ、（ｂ）が２ｍの長さの振り子を制御対象として用いた場合のグラフである。

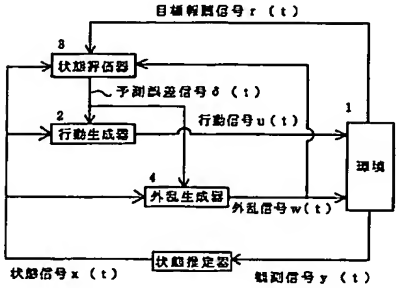
＊【図６】第２具体例の概略図である。

【図７】第２具体例の直動アクチュエータの位置変化のグラフで、（ａ）が１ｋｇの質量の搬送物を用いた場合のグラフ、（ｂ）が３ｋｇの質量の搬送物を用いた場合のグラフである。

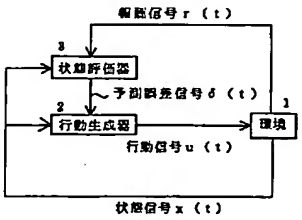
【符号の説明】

ｑ(t)　評価信号
ｒ(t)　目標報酬信号
ｓ(t)　外乱報酬信号
ｕ(t)　行動信号
ｗ(t)　外乱信号
１　環境
２　行動生成器
３　状態評価器
４　外乱生成器

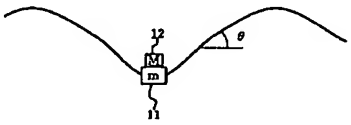
【図１】



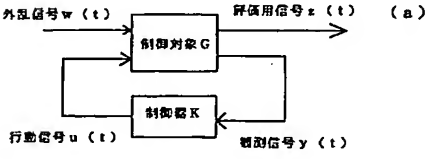
【図２】



【図６】

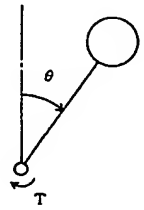


【図３】

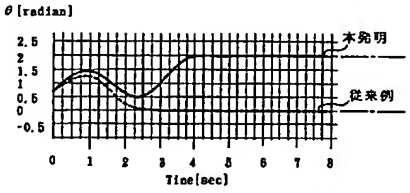


【図４】

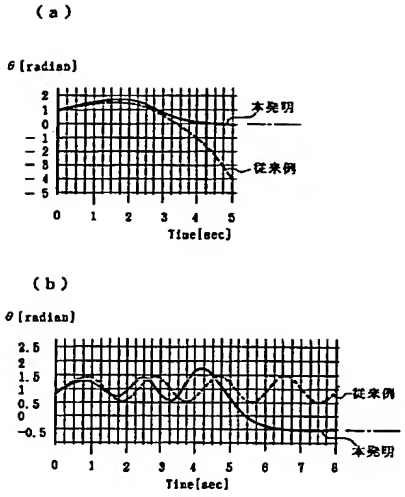
（ａ）



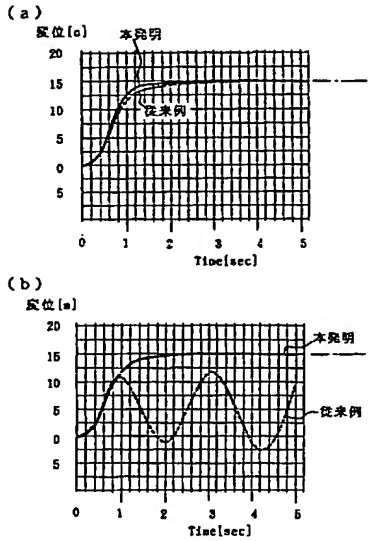
（ｂ）



【図５】



【図７】



フロントページの続き

(72)発明者　銅谷　賢治
京都府相楽郡精華町光台　７－２－１－５－２０１

Ｆターム(参考)　5H004 GA07 GA15 GA17 JA13 JB22 KC09 KC18 KC28 KD42 KD62